GETTING STARTED

This Python script is designed to extract data from Non-Conformance Reports (NCRs) stored in PDF files and export the information into an Excel file. The script can process individual PDF files or a folder containing multiple PDF files. It extracts relevant data from the PDF tables and organizes it in an Excel sheet in a desired format.

➢ Necessary libraries to be installed (pip install requirement.txt)

1) tabula-py: Python library to extract tables from PDF Files. You can install it using the command **: pip install tabula-py**
2) pandas: This library is used for data manipulation and analysis. You can install it using the command: **pip install pandas**
3) openpyxl (Python library for working with Excel files). You can install it using the command: **pip install openpyxl**

➢ How to run the script

1) Clone or download this repository to your local machine.
2) Open a terminal or command prompt and navigate to the directory containing the script and input files (if applicable).
3) If your input is a single file make sure to change the variable file_or_folder == 2: to file_or_folder == 1: in line 480. Else leave the script as it is.
4) Run the script using Python, and it will ask you the following.
5) Folder Input: It will ask the following

   "Enter the folder path (Hit enter if input is a file path):" Enter the path of the folder containing all the files here. If the input is just one file simply press enter.

```
(project) PS C:\Local\Project> & c:/tools/pyenvs/project/Scripts/python.exe c:/Local/Project/NCR-Extraction-Version1.1.0.py
Enter the folder path(Hit enter if input is a file path): C:\Local\Project\NonConformanceReport\SampleInputFolder
Enter the file path(Hit enter if input is a folder path):
['C:\\Local\\Project\\NonConformanceReport\\SampleInputFolder\\NCR -Dummy 3.pdf', 'C:\\Local\\Project\\NonConformanceReport\\SampleInputFolder\\NCR-dummy.pdf', 'C:\\Local\\Project\\NonConformanceReport\\Sa
mpleInputFolder\\NCR-Dummy2 .pdf', 'C:\\Local\\Project\\NonConformanceReport\\SampleInputFolder\\NCR_1.pdf']
Performig operation on file: C:\Local\Project\NonConformanceReport\SampleInputFolder\NCR -Dummy 3.pdf
Performig operation on file: C:\Local\Project\NonConformanceReport\SampleInputFolder\NCR-dummy.pdf
Performig operation on file: C:\Local\Project\NonConformanceReport\SampleInputFolder\NCR-Dummy2 .pdf
Performig operation on file: C:\Local\Project\NonConformanceReport\SampleInputFolder\NCR_1.pdf
(project) PS C:\Local\Project>
```

6) File Input: It will ask the following

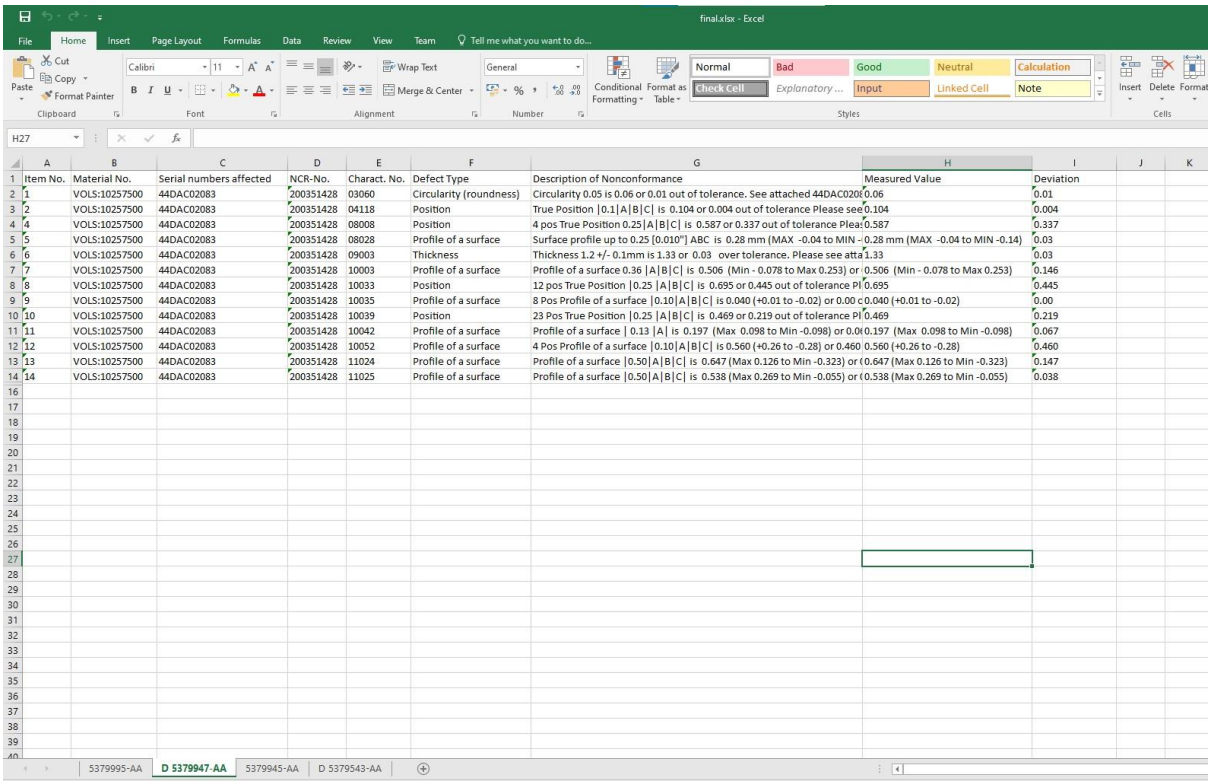
   "Enter the file path (Hit enter if input is a folder path):" Enter the path of the file. If you have already given an input as a folder simply press enter.

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL

(project) PS C:\Local\Project> & c:/tools/pyenvs/project/Scripts/python.exe c:/Local/Project/NCR-Extraction-Version1.1.0.py
Enter the folder path(Hit enter if input is a file path):
Enter the file path(Hit enter if input is a folder path): C:\Local\Project\NonConformanceReport\SampleInputFile\NCR final.pdf
```

7) It will start execution.
8) After execution, the script will create an Excel file named "final.xlsx" in the same directory, containing the extracted data.

   Sample Output:



Item No.	Material No.	Serial numbers affected	NCR-No.	Charact. No.	Defect Type	Description of Nonconformance	Measured Value	Deviation
1	VOLS:10257500	44DAC02083	200351428	03060	Circularity (roundness)	Circularity 0.05 is 0.06 or 0.01 out of tolerance. See attached 44DAC0208	0.06	0.01
2	VOLS:10257500	44DAC02083	200351428	04118	Position	True Position |0.1|A|B|C| is 0.104 or 0.004 out of tolerance Please see	0.104	0.004
4	VOLS:10257500	44DAC02083	200351428	08008	Position	4 pos True Position 0.25|A|B|C| is 0.587 or 0.337 out of tolerance Plea	0.587	0.337
5	VOLS:10257500	44DAC02083	200351428	08028	Profile of a surface	Surface profile up to 0.25 [0.010"] ABC is 0.28 mm (MAX -0.04 to MIN -	0.28 mm (MAX -0.04 to MIN -0.14)	0.03
6	VOLS:10257500	44DAC02083	200351428	09003	Thickness	Thickness 1.2 +/- 0.1mm is 1.33 or 0.03 over tolerance. Please see atta	1.33	0.03
7	VOLS:10257500	44DAC02083	200351428	10003	Profile of a surface	Profile of a surface 0.36 |A|B|C| is 0.506 (Min - 0.078 to Max 0.253) or	0.506 (Min - 0.078 to Max 0.253)	0.146
8	VOLS:10257500	44DAC02083	200351428	10033	Position	12 pos True Position |0.25 |A|B|C| is 0.695 or 0.445 out of tolerance Pl	0.695	0.445
9	VOLS:10257500	44DAC02083	200351428	10035	Profile of a surface	8 Pos Profile of a surface |0.10|A|B|C| is 0.040 (+0.01 to -0.02) or 0.00 c	0.040 (+0.01 to -0.02)	0.00
10	VOLS:10257500	44DAC02083	200351428	10039	Position	23 Pos True Position |0.25 |A|B|C| is 0.469 or 0.219 out of tolerance Pl	0.469	0.219
11	VOLS:10257500	44DAC02083	200351428	10042	Profile of a surface	Profile of a surface | 0.13 |A| is 0.197 (Max 0.098 to Min -0.098) or 0.0	0.197 (Max 0.098 to Min -0.098)	0.067
12	VOLS:10257500	44DAC02083	200351428	10052	Profile of a surface	4 Pos Profile of a surface |0.10|A|B|C| is 0.560 (+0.26 to -0.28) or 0.460	0.560 (+0.26 to -0.28)	0.460
13	VOLS:10257500	44DAC02083	200351428	11024	Profile of a surface	Profile of a surface |0.50|A|B|C| is 0.647 (Max 0.126 to Min -0.323) or (	0.647 (Max 0.126 to Min -0.323)	0.147
14	VOLS:10257500	44DAC02083	200351428	11025	Profile of a surface	Profile of a surface |0.50|A|B|C| is 0.538 (Max 0.269 to Min -0.055) or (	0.538 (Max 0.269 to Min -0.055)	0.038

➢ Code Explanation

There are multiple functions in order to perform the desired task.

1. folder_list(folder_path): This function takes a folder path as input and returns a list of PDF files found in that folder.
2. valid_sheet(string): A utility function to remove invalid characters from a string to be used as a valid sheet name in Excel.
3. not_null_input(input_value): A utility function to check if an input value is not null and return its string representation.
4. read_pdf(pdf_path, excel_path): Reads tables from a PDF file and saves them to an Excel file.
5. read_pdf_page1_table1(workbook): Extracts data from the first page of the PDF table and returns it as a list.
6. find_last_row(file_path, sheet_name, column_index): Finds the last filled row in an Excel sheet.
7. find_last_row2(workbook, column_index, sheet_name): Similar to find_last_row but accepts a workbook object.
8. read_pdf_page1_table2(workbook): Reads data from the second table on the first page of the PDF and returns it as a list.
9. read_pdf_page1_table100(workbook): Reads data from the 100th table on the first page of the PDF and returns it as a list.
10. read_pdf_tables_3_4_5(pdf_file): Extracts specific data from the PDF using regular expressions and returns the results as lists.
11. write_to_excel(string, row_no, col_no, file_path, sheet_name): Writes a string value to a specific cell in an Excel sheet.
12. setup_output_excel(sheet_name, output_path): Sets up the output Excel sheet with header values if it's a new sheet.
13. print_group_1(table1, start, count, output_path, sheet_name): Writes data from table1 to the Excel sheet.
14. print_group_2(table2, start, count, output_path, sheet_name): Writes data from table2 to the Excel sheet.

15. print_group_100(table100, start, count, output_path, sheet_name): Writes data from table100 to the Excel sheet.

16. print_group_3_4_5(table3, table4, table5, start_point, count, output_path, sheet_name): Writes data from table3, table4, and table5 to the Excel sheet.

17. print_group_6_7(table6, table7, start_point, count, output_path, sheet_name): Writes data from table6 and table7 to the Excel sheet.

18. get_dev_and_mes(sentences): Extracts measured and deviation values from a list of sentences.

19. perform_all_file_actions(pdf_path, output_path): Executes all the file processing steps for a single PDF file.

➢ Important Notes

1) The script may require adjustments if the PDFs do not conform to the expected format or contain different data patterns. Also, make sure you have the required libraries (tabula, pandas, openpyxl) installed before running the script.

2) If the PDFs have different table structures or contain variations, additional adjustments to the script may be necessary.